W510–W515 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* *Published online 4 June 2012*
doi:10.1093/nar/gks483

Cscan: finding common regulators of a set of genes by using a collection of genome-wide ChIP-seq datasets

Federico Zambelli[1], Gian Marco Prazzoli[1], Graziano Pesole[2,3] and Giulio Pavesi[1,*]

[1]Dipartimento di Scienze Biomolecolari e Biotecnologie, Università di Milano, Italy, [2]Istituto di Biomembrane e Bioenergetica, Consiglio Nazionale delle Ricerche, Bari, Italy and [3]Dipartimento di Bioscienze, Biotecnologie e Scienze Farmacologiche, Università di Bari, Italy

Received February 13, 2012; Revised April 25, 2012; Accepted May 2, 2012

ABSTRACT

The regulation of transcription of eukaryotic genes is a very complex process, which involves interactions between transcription factors (TFs) and DNA, as well as other epigenetic factors like histone modifications, DNA methylation, and so on, which nowadays can be studied and characterized with techniques like ChIP-Seq. Cscan is a web resource that includes a large collection of genome-wide ChIP-Seq experiments performed on TFs, histone modifications, RNA polymerases and others. Enriched peak regions from the ChIP-Seq experiments are crossed with the genomic coordinates of a set of input genes, to identify which of the experiments present a statistically significant number of peaks within the input genes' loci. The input can be a cluster of co-expressed genes, or any other set of genes sharing a common regulatory profile. Users can thus single out which TFs are likely to be common regulators of the genes, and their respective correlations. Also, by examining results on promoter activation, transcription, histone modifications, polymerase binding and so on, users can investigate the effect of the TFs (activation or repression of transcription) as well as of the cell or tissue specificity of the genes' regulation and expression. The web interface is free for use, and there is no login requirement. Available at: http://www.beaconlab.it/cscan.

INTRODUCTION

The regulation of eukaryotic gene transcription is a very complex process, which depends on interactions between transcription factors (TFs) and DNA, as well as on chromatin structure and other epigenetic factors such as histone modifications, DNA methylation and so on. Research in this field has witnessed a major leap forward with the introduction of techniques like Chromatin Immunoprecipitation (ChIP) (1), which, followed by the employment of genome tiling oligonucleotide arrays [ChIP on Chip (2)] or next-generation sequencing [ChIP-Seq (3)], permits to build genome-wide maps of TF binding, histone modifications or any other DNA interacting protein involved in transcription regulation. ChIP-Seq has rapidly become the method of choice for research in this field, given its higher resolution with respect to ChIP on Chip, and the constantly decreasing cost of next-generation sequencing experiments. As a consequence, today genomic resources like the UCSC Genome Browser (4) or dedicated databases like hmChip (5) make available for retrieval the genomic maps of hundreds of TFs, as well as of histone modifications, PolII and PolIII binding, and so on, in several different cell lines. Thus, starting from a gene, its putative regulators as well as epigenetic information associated with it can be easily retrieved vice versa, different ChIP-Seq experiments can be correlated with one another by comparing the distribution of the sequence reads of each one (6) and, once the list of genomic binding regions for a TF is available, the target genes it is likely to regulate can be easily singled out by using tools like GREAT (7).

On the other hand, a very common problem that researchers have to face is, given a set of genes showing similar expression patterns, to find out which common regulators they share, responsible for the expression observed. This type of analysis is often performed by finding similar and over-represented sequence elements, for example in promoter sequences, either by using de novo motif finding tools (8) or descriptors of the binding specificity of TFs (9,10). While useful in many cases, these approaches suffer from several limitations: the binding

*To whom correspondence should be addressed. Tel: +39 50314884; Fax: +39 50315044; Email: giulio.pavesi@unimi.it

specificity of many TFs is as yet unknown or not well characterized; different TFs have very similar binding sites, making difficult, given a sequence motif, to assess which TF actually could bind it; some key regulators do not bind DNA directly, but act as co-factors with TFs; sequence motifs are often weakly conserved, and hard to discriminate against random similarities; sequence analysis tools usually ignore chromatin structure and DNA accessibility, usually resulting in an 'over-prediction' of sequence motifs.

The web tool we present, named Cscan, is based on a large collection of ChIP-Seq experiments for several TFs and other factors related to transcription regulation. Enriched regions from the ChIP-Seq experiments have been crossed with the genomic coordinates of available RefSeq and Ensembl gene annotations, so to build genome-wide maps of putative target genes in each experiment. Given a set of genes as input, the interface evaluates the over- (or under-) representation of target sites for the DNA binding protein considered in each ChIP experiment by counting the number of target genes in the experiment contained in the input set, and comparing this count to the overall genome-wide number of its target genes to assess statistical significance with a Fisher's exact test. Experiments with a significantly high number of sites within the input genes' loci are thus likely to correspond to TFs, which are common regulators of the genes. The computation is performed for hundreds of different TFs with other data like histone modifications and RNA polymerases (and/or their subunits), so to provide a more comprehensive view of all the genetic and epigenetic factors involved in the regulation of the input genes, and their effect on gene transcription.

ChIP-SEQ DATA COLLECTION

We retrieved the ChIP-Seq peak lists publicly available and already past the public release date at the UCSC Genome Browser for TFs, histone modifications, and RNA polymerases produced by the ENCODE project (11). Also, we retrieved from the original publications the datasets that have been included in the hmChip database (5). Finally, we added the HMMChip tracks of the UCSC Genome Browser, showing chromatin state segmentation for each of nine human cell types. A common set of states (including for example active promoter, weak promoter, repressed transcription, and so on) across the cell types were annotated integrating ChIP-Seq data for nine histone modifications using a Hidden Markov Model. The genome was thus segmented into regions according to the corresponding chromatin state (12).

Overall, data collection resulted for human in 409 different experiments for 144 TFs or co-factors in 65 different cell lines, 234 experiments for 11 different histone modifications in 23 cell lines, 46 experiments for 6 RNA polymerases (or their subunits) in 28 cell lines, data for CTCF binding in 49 cell lines, for a total of 777 different experiments or annotations in 102 cell lines. We are currently populating the mouse collection, which as of today contains data for about 50 TFs.

In each ChIP-Seq dataset, the genomic coordinates of each region marked as 'peak' have been crossed with the RefSeq or Ensembl gene annotations available. This resulted in a table with one row for each annotated gene, and one column for each ChIP-Seq experiment. The table reports the presence/absence of a peak in the ChIP-Seq experiment within different regions of the locus of the gene (i.e. in its promoter/upstream of the TSS at different distances, within the transcribed region, and so on, see Supplementary Figure S1).

FINDING COMMON REGULATORS

Starting from the data collected, let G be a sample of k genes or transcripts. If a given TF is a common regulator of the genes, then one should find an enrichment of binding regions for the TF associated with the genes, e.g. in their promoters or transcribed regions. For example, let m be the number of genes in the sample that have a peak for the TF in their promoter. Then, let N be the number of annotated genes in the genome and let n be the number of annotated genes in the genome that contain a ChIP-Seq peak for the TF in their promoter. The enrichment of the TF binding sites with respect to the gene sample can be thus evaluated by using a Fisher's exact test (hypergeometric distribution) with N, n, k and m as parameters.

The same principle can be applied to any other type of genome-wide ChIP experiment. For example, we can assess whether a given histone modification can be associated with the genes' promoters, hence denoting e.g. if their promoters as active or repressed, or whether RNA polymerase binding in a given cell line is enriched in the set of genes denoting their transcription, and so on.

THE USER INTERFACE

The user interface contains two main panels on the left and right hand side, which can be used to input a set of genes for finding their common regulators or for browsing and retrieving data from the ChIP-Seq data collections available.

Data browsing and retrieval

The right hand panel allows users to browse the data Cscan is based on, and to retrieve the list of target genes associated with a given experiment of interest. Users can select (i) the protein that has been ChIP'ed (ii) the cell line in which the experiment was performed, (iii) the region of the genes' locus in which peaks have to be located for the gene to be considered as a target (e.g. the −450, +50 region including the core promoter or the transcribed region including 1 kbp upstream), (iv) the source organism and the gene annotation to be used to display the results (RefSeq or Ensembl) and (v) the genome assembly used in the study. Once any of the input fields is selected, the other choices are automatically limited to the experiments available, e.g. once a TF has been selected in the list, the selection of the cell lines will be limited to those for which data are available for the TF, and so on.

The output will be displayed within an 'Experiment view' output window, described in the 'Output' section. Alternatively, given a gene (transcript) identifier, users can retrieve the list of ChIP-Seq experiments in the database that present a peak within the gene region defined.

Gene input

The left hand panel is used to input a set of genes, by using the RefSeq or Ensembl IDs of their respective transcripts, and finding ChIP-Seq experiments that have a significantly high (or low) number of peaks associated with the genes. Users then have to specify the following: (i) the source organism of the genes (at the present time, human or mouse); (ii) the region, with respect to the gene, that has to be analyzed (e.g. core promoter only or upstream and transcribed regions); and (iii) the cell line in which the ChIP-Seq data used for the analysis were generated (or this parameter can be set to 'ALL' indicating that all the data available have to be used). A typical analysis takes a few seconds, and results will appear in the middle of the page.

Output

The output is split into two tables as shown in Figure 1a. The topmost one is dedicated to TFs (or co-factors), while the bottom one contains results for CTCF, histone modifications, PolII and PolIII binding, HMMChip regions, and other experiments not involving TFs (denoted as 'Features' in the table). A link on the top of the features table gives further explanations on each one, and its possible effect on the regulation of the genes. In each table, the ChIP-Seq experiments used in the analysis are ranked according to the *P*-value of the Fisher's test. From left to right the columns of the output table summarize the following:

- [TF/FEATURE]: The TF or feature of the ChIP-Seq experiment;
- [LINE]: The cell line in which the experiment was performed;

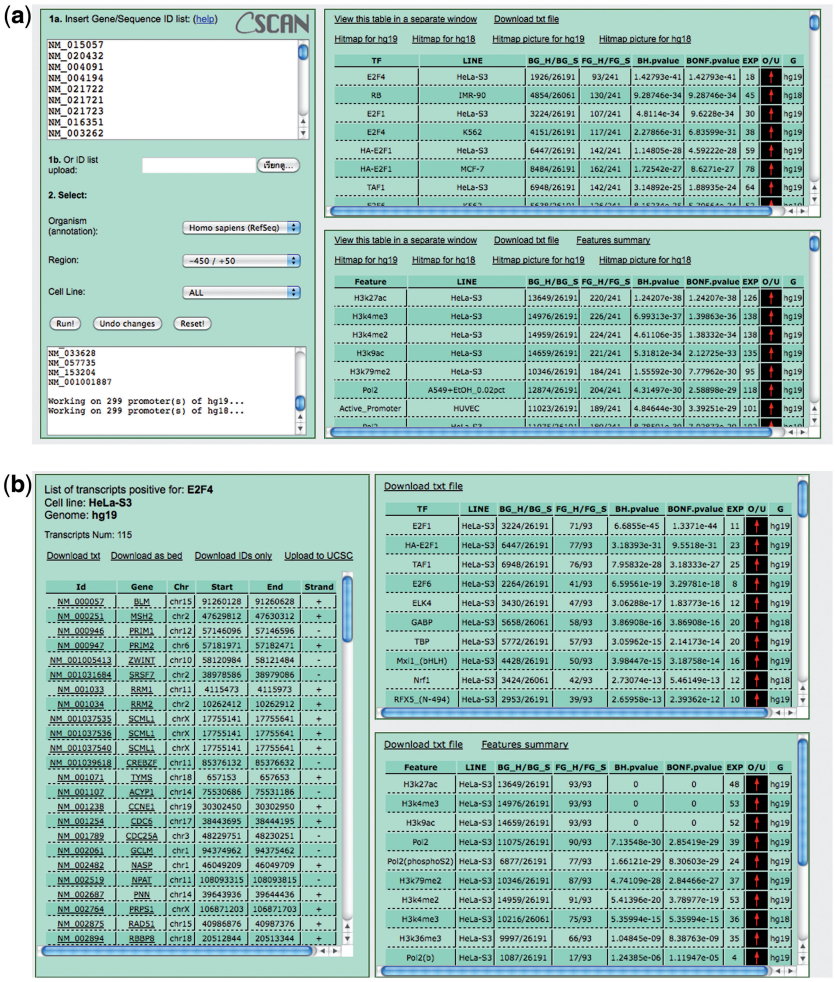

Figure 1. (**a**) Example of the output of Cscan showing the list of input genes (input box on the left hand side) as well as the TFs list (right, top) and 'features' (histone modifications, polymerase binding and so on, bottom right) ranked according to the resulting *P* value. See the main text for further explanation on the output fields. (**b**) The 'Experiment view' table, showing for a TF (E2F4) in the selected cell line (HeLa-S3) the target genes that were included in the input sample (left). The tables on the right show enrichment of other TFs (top) and features (bottom), computed on the E2F4 target genes.

- [BG_H/BG_S]: The number BG_H of genes in the genome annotation which contain a peak for the experiment in the region selected as input, and the overall number BG_S of 'background' genes (e.g. in the whole genome annotation used);
- [FG_H/FG_S]: As in the previous point, but restricted to the FG_H genes that contain a peak out of the total FG_S input genes;
- [Bonf-Pvalue]: The Bonferroni corrected *P* value computed with the Fisher's test (hypergeometric distribution) according to the BG_S, BG_H, FG_S and FG_H values;
- [BH-Pvalue]: The Benjamini-Hochberg corrected *P* value. This correction yields less restrictive *P* values than the Bonferroni one. Users can choose which of the two seems to be more suitable for their data;
- [EXP]: The expected value for FG_H, according to FG_S, BG_H and BG_S; and
- [O/U]: A red arrow pointing upwards if the number of FG_H genes is greater than the expected value, a green arrow pointing downwards if lower. The arrow thus denotes whether peaks for the ChIP-Seq experiment are under- or over-represented in the gene sample provided as input.

Experiments that present a large number of genome-wide targets (i.e. more than one third of the annotated transcripts), thus unlikely to provide meaningful information, have the corresponding line shaded in grey. Clicking on the links above each table allows users to open the table in a new window (discussed later), to download the table in text format, or to display the 'hitmap', which associates with each input gene and ChIP experiment pair a '1' if the gene region specified as input contains at least one ChIP peak, '0' if not, and that can be used for further computations and analyses. The hitmap can be displayed also as a picture, with a colored spot in correspondence of experiment peak-gene associations, black otherwise.

Once an output table is opened in a new window (called 'TF table view' and 'Feature view', respectively), additional links appear. Clicking on the 'info' icons (a white 'i' on a blue background) provides further information (if available) on the subject of the ChIP experiment or the cell line in which it was performed. Clicking on the TF/feature name, instead, opens the 'Experiment view' window, that displays the list of input genes which are associated with a peak for the ChIP-Seq experiment and cell line selected, as well as their genomic coordinates. From this window, users can download the list of gene IDs, or the .bed file of their genomic coordinates which can be uploaded automatically to the UCSC Genome Browser for further analysis. The 'Get Correlations' button on the right-hand side performs another run of Cscan, but restricted to the list of target genes for the TF/feature currently investigated, and using experiments performed in the same cell line: in this way, users can immediately assess which other experiments have significant correlation (or anti-correlation) with the TF/feature on the set of genes studied (Figure 1b).

This 'Experiment view' window is also displayed when a given experiment is selected by using the right-hand panel of the interface: in this case, the list of targets will comprise all the target genes in the genome annotation available.

EXAMPLES

As mentioned before, Cscan can be applied to different types of analysis. A straightforward application is to study clusters of genes with similar expression patterns, so to single out their putative common regulators. But, as epigenetic data are also included in the analysis, by crossing these data with the results on TF binding one can get an idea also on the effect of the TF regulation (activation/repression) and/or on the tissue/cell/condition specificity of TF binding. Also, if a novel ChIP-Seq experiment has been performed, Cscan allows for an immediate assessment of which other TFs show significance correlation or anti-correlation with the studied one, as well as of whether the TF correlates with histone modifications, active/repressed promoters, or polymerase binding, the latter indicating again whether it might act as an activator or a repressor. Finally, the results of Cscan provide an immediate validation for predictions derived from other tools, for example conserved motifs found by sequence analysis and motif discovery methods. Users can thus submit the same set of genes simultaneously to Cscan and to tools like Pscan (9) or Clover (10), and assess whether TFs singled out by sequence analysis are also detected by Cscan, or which TFs are more likely to bind a given sequence motif. These two approaches can also be seen as complementary, because ChIP data are not available for all the TFs and vice versa, a reliable binding descriptor is not available for all the TFs.

In the following section, we describe some examples of usage of Cscan. The corresponding datasets are included in the interface and can be easily loaded as input by clicking on the corresponding link. The analyses were performed by using as a target region of the genes the core promoter (from −450 to +50 with respect to the transcription start site). The results are also provided in Supplementary Table S1.

Human cell cycle regulated genes

We retrieved the clusters of human genes whose expression has been characterized of being specific of a given phase of the cell cycle [G1/S, S, G2, G2/M and M/G1 (13)]. The microarray experiment was performed in HeLa cells.

Concerning the 'features' table, nearly all the genes of each dataset seem to be transcribed in all the cell lines available, and not only HeLa cells. Indeed, PolII, 'Active promoter' HMMChIP annotations, and histone modifications associated with active promoters and transcription are highly enriched, while those features associated with gene silencing are significantly under represented. This is hardly a surprising result, because we can expect cell-cycle expressed genes not to be cell line or tissue specific. Figure 2 and Supplementary Table 1 summarize the most significant TFs in the five phases (Bonferroni corrected $P < 10^{-5}$ in at least one).

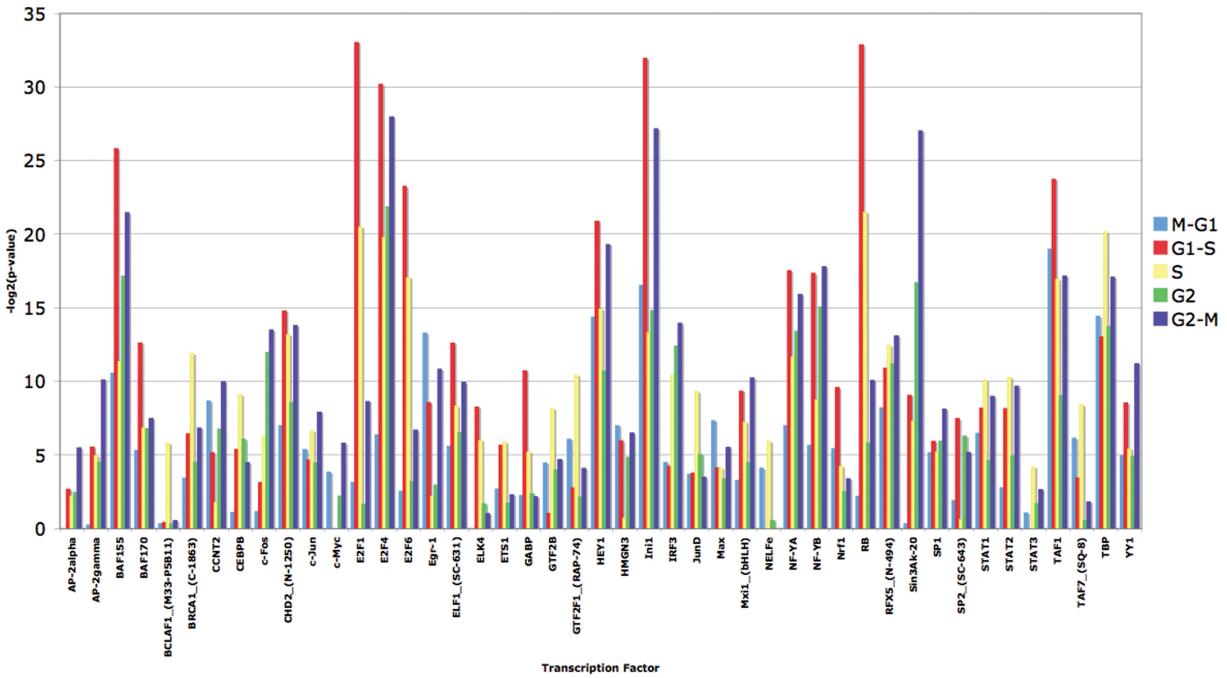

Figure 2. The most significantly enriched TFs in the different phases of human cell cycle ($P < 10^{-5}$ in at least one set). We considered experiments performed on the same cell line of the expression data (HeLa). For TFs for which HeLa data are not yet available, we employed K562 data.

TFs of the list showing highest enrichments are known cell cycle regulators, and as expected their over-representation in the input sets changes according to the different phases. For example, all members of the E2F family with available data are significantly enriched in phases G1/S and S. E2F1 and E2F6, however, drop in the successive phases, while E2F4 remains enriched throughout G2 and G2/M. The Retinoblastoma protein (RB) lacks a DNA binding domain and is recruited to promoters by other sequence-specific TFs, such as the members of the E2F family: indeed its enrichment profile shows similarity with E2F1 and E2F6. Thus, results of this kind would be hard to obtain with sequence analysis alone, also because specialized databases like TRANSFAC and JASPAR report only a generic 'E2F' binding motif, while for example E2F4 has been recently shown to bind the CHR promoter element as part of the DREAM complex (14).

Human tissue-specific genes

As an example of analysis of tissue-specific genes, we built two datasets of liver and whole brain-specific genes using the Gene Sorter tool at the UCSC Genome Browser. We selected RefSeq genes with an expression logBase2 (tissue/reference) value >2 in the two tissues.

The 'features' results on liver genes show how their transcription activation, active promoter marks and PolII binding seem to be confined to HepG2 cells, which indeed are hepatocarcinoma cells, and a model system for the study of polarized human hepatocytes. On the other hand, the signatures associated to transcription repression and gene silencing are over-represented in all other cell lines. Also, the TF table shows as significantly enriched a series of TFs (HNF4A, HNF4G, RXRA, FOXA1, and so on) known to be associated with tissue-specific gene expression in liver or liver development. Other TFs usually associated with cell cycle or 'housekeeping' gene expression do not show any enrichment on this gene set in HepG2 cells. However, not all the genes of this set are associated with PolII binding or active promoters and TF binding. This fact can be due to different reasons, like experimental issues (false positives in the microarray experiment or false negatives in the ChIP-Seq analyses producing the lists of peaks), or to differences between normal and tumoral liver cells. Another possibility is that, as multiple promoters can be associated with the same gene, Cscan is able to mark which promoter is active and bound by TFs in the cell line investigated.

The result on 'brain-specific' genes shows how they do not seem to be transcribed in any of the cell lines available, nor are enriched for any histone mark associated with transcriptional activation. The TF table likewise shows how virtually all the TFs are underrepresented in the gene sample, with the sole exception of NRSF (Neuron-restrictive Silencer transcription Factor) throughout different cell lines, which indeed is a repressor protein expressed in non-neuronal tissues, repressing the expression of several neuronal genes.

Computing correlations between different ChIP-Seq experiments

A simple but explicative example on how Cscan can be used to identify correlations among different ChIP-Seq experiments is the set of target promoters for BDP1 (B double-prime 1) in human HeLa cells, retrievable from Cscan itself. BDP1 is a subunit of the TFIIIB

transcription initiation complex, which recruits RNA polymerase III to target promoters to activate its transcription (15). Indeed, the features table shows that PolIII associated with the promoters of the genes. In the TF list, the highest correlations are with BDP1 itself in a different cell line (K562), showing how BDF1 binding does not seem to be cell-line specific. Also, all the targets of another factor (BRF1) are included into the BDF1 list. Indeed, BDF1 is another subunit of the same complex, together with TFIIIC-110, which is also highly enriched. Other regulators, not related to PolIII transcription appear anyway to be over-represented. Although, for example TATA-binding protein has already been shown to be a regulator of PolIII transcribed genes, other factors like STAT1 in interferon-stimulated cells or heat-shock protein that target most of the genes show how they are probably activated and involved in several different pathways.

CONCLUSIONS

Cscan is a web server that employs a collection of several hundreds of different ChIP-Seq experiments to identify putative common regulators in a set of genes, as well as assessing their transcriptional and epigenetic profile. Clearly, results depend on the presence of a given TF or cell line in the collection of experiments the server is based on, and while for example we have already a good coverage for tissues like liver we still lack data on tissue-specific TFs and epigenomic information in several tissues or cell lines. We can expect however this gap to be quickly filled in the near future, given the ever increasing amount of ChIP-Seq experiments that are performed and published almost on a daily basis. Also, we plan in the near future to include information about distal regulatory elements in enhancers/silencers, by crossing data on TF binding with chromatin signatures marking likely enhancer regions and with CTCF binding to insulators, so to overcome the obvious limitations of analyses only on promoters or transcribed regions. Concerning the extension of Cscan to other species, we are currently populating the ChIP-Seq mouse data collection, as well as preparing the inclusion of other species, from yeast to the data produced by the modENCODE project on *Caenorhabditis elegans* and *Drosophila* (16).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1 and Supplementary Figure 1.

FUNDING

Italian Ministry of University and Research Fondo Italiano per la Ricerca di Base (FIRB) project 'Laboratorio Internazionale di Bioinformatica' (LIBI); Consiglio Nazionale delle Ricerche (CNR) flagship project EPIGEN. Funding for open access charge: Italian Ministry of University and Research Fondo Italiano per la Ricerca di Base (FIRB) project 'Laboratorio Internazionale di Bioinformatica' (LIBI).

Conflict of interest statement. None declared.

REFERENCES

1. Collas,P. and Dahl,J.A. (2008) Chop it, ChIP it, check it: the current status of chromatin immunoprecipitation. *Front. Biosci.*, **13**, 929–943.
2. Pillai,S. and Chellappan,S.P. (2009) ChIP on chip assays: genome-wide analysis of transcription factor binding and histone modifications. *Methods Mol. Biol.*, **523**, 341–366.
3. Mardis,E.R. (2007) ChIP-seq: welcome to the new frontier. *Nat. Methods*, **4**, 613–614.
4. Fujita,P.A., Rhead,B., Zweig,A.S., Hinrichs,A.S., Karolchik,D., Cline,M.S., Goldman,M., Barber,G.P., Clawson,H., Coelho,A. *et al.* (2011) The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.*, **39**, D876–D882.
5. Chen,L., Wu,G. and Ji,H. (2010) hmChIP: a database and web server for exploring publicly available human and mouse ChIP-seq and ChIP-chip data. *Bioinformatics*, **27**, 1447–1448.
6. Ye,T., Krebs,A.R., Choukrallah,M.A., Keime,C., Plewniak,F., Davidson,I. and Tora,L. (2011) seqMINER: an integrated ChIP-seq data interpretation platform. *Nucleic Acids Res.*, **39**, e35.
7. McLean,C.Y., Bristor,D., Hiller,M., Clarke,S.L., Schaar,B.T., Lowe,C.B., Wenger,A.M. and Bejerano,G. (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, **28**, 495–501.
8. Pavesi,G., Mauri,G. and Pesole,G. (2004) In silico representation and discovery of transcription factor binding sites. *Brief Bioinform.*, **5**, 217–236.
9. Zambelli,F., Pesole,G. and Pavesi,G. (2009) Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. *Nucleic Acids Res.*, **37**, W247–W252.
10. Frith,M.C., Fu,Y., Yu,L., Chen,J.F., Hansen,U. and Weng,Z. (2004) Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res.*, **32**, 1372–1381.
11. Rosenbloom,K.R., Dreszer,T.R., Long,J.C., Malladi,V.S., Sloan,C.A., Raney,B.J., Cline,M.S., Karolchik,D., Barber,G.P., Clawson,H. *et al.* (2012) ENCODE whole-genome data in the UCSC Genome Browser: update 2012. *Nucleic Acids Res.*, **40**, D912–D917.
12. Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.
13. Whitfield,M.L., Sherlock,G., Saldanha,A.J., Murray,J.I., Ball,C.A., Alexander,K.E., Matese,J.C., Perou,C.M., Hurt,M.M., Brown,P.O. *et al.* (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell.*, **13**, 1977–2000.
14. Muller,G.A., Quaas,M., Schumann,M., Krause,E., Padi,M., Fischer,M., Litovchick,L., Decaprio,J.A. and Engeland,K. (2012) The CHR promoter element controls cell cycle-dependent gene transcription and binds the DREAM and MMB complexes. *Nucleic Acids Res.*, **40**, 1561–1578.
15. Noma,K. and Kamakaka,R.T. (2010) The human Pol III transcriptome and gene information flow. *Nat. Struct. Mol. Biol.*, **17**, 539–541.
16. Contrino,S., Smith,R.N., Butano,D., Carr,A., Hu,F., Lyne,R., Rutherford,K., Kalderimis,A., Sullivan,J., Carbon,S. *et al.* (2012) modMine: flexible access to modENCODE data. *Nucleic Acids Res.*, **40**, D1082–D1088.